DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 7

11-1-2007

The Non-Parametric Difference Score: A Workable Solution for Analyzing Two-Wave Change When The Measures Themselves Change Across Waves

Jennifer E. V. Lloyd
University of British Columbia

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Recommended Citation

Lloyd, Jennifer E. V. and Zumbo, Bruno D. (2007) "The Non-Parametric Difference Score: A Workable Solution for Analyzing Two-Wave Change When The Measures Themselves Change Across Waves," *Journal of Modern Applied Statistical Methods*: Vol. 6 : Iss. 2 , Article 7.
DOI: 10.22237/jmasm/1193889960
Available at: http://digitalcommons.wayne.edu/jmasm/vol6/iss2/7

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 413- 420

1538 – 9472/07/$95.00

The Non-Parametric Difference Score: A Workable Solution for Analyzing Two-Wave Change When The Measures Themselves Change Across Waves

Jennifer E. V. Lloyd Bruno D. Zumbo
University of British Columbia

The non-parametric difference score is introduced. It is a workable solution to the problem of analyzing change over two waves (i.e., a pretest-posttest design) when the measures themselves vary over time. An example highlighting the solution's implementation is provided, as is a discussion of the solution's assumptions, strengths, and limitations.

Key words: Non-parametric, difference score, two-wave, change, quantitative analysis.

Introduction

Individual change is the subject of significant attention in education, health, and the social sciences. The analysis of such change is aimed at quantifying the amount by which individuals grow, mature, improve, and progress over time. By measuring and tracking changes, it is possible to reveal the temporal nature of development (Singer & Willett, 2003).

This temporal nature of development may be studied over varied spans of time: hours, days, weeks, months, and even years. Waves are the measurement occasions or periods of data

This temporal nature of development may be studied over varied spans of time: hours, days, weeks, months, and even years. Waves are the measurement occasions or periods of data

collection that are plan-fully interspersed throughout these spans of time. Two-wave designs, often known as pretest-posttest designs, are the specific focus of this article. Such designs allow for relatively straightforward appraisal of a treatment effect by detecting differences in a given outcome across two waves – typically before the treatment and after it. Such differences normally represent the comparison of test-takers' scores at the second wave of data collection to their respective baseline or initial measure scores (Zumbo, 1999). Lloyd (2006) and Lloyd, Zumbo, and Siegel (2007) explore the problem of analyzing change and growth when the measures themselves change across multiple (i.e., three or more) waves.

Repeated Measures Analyses: Three Research Scenarios

Several familiar parametric methodologies, called repeated measures analyses, centre upon quantifying change over time. As described by Lloyd (2006) and Lloyd, Zumbo, and Siegel (2006), these methodologies are generally used in three research scenarios:

Scenario 1: Exact same measure across both waves

In this scenario, one's construct of choice makes possible the use and re-use of the

Jennifer E.V. Lloyd is a Research Scientist and Post-Doctoral Fellow for the Human Early Learning Partnership at the University of British Columbia (UBC). Email her at jennifer.lloyd@ubc.ca. Bruno D. Zumbo is Professor, Measurement, Evaluation, and Research Methodology (MERM) Program, and Department of Statistics and the Institute of Applied Mathematics at the University of British Columbia. Email him at bruno.zumbo@ubc.ca. The authors thank the Social Sciences and Humanities Research Council of Canada for funding this research project, and the British Columbia Ministry of Education for approving the use of their data. Thanks are also extended to Dr. Anita Hubley, Dr. Kimberly Schonert-Reichl, and Alex Mann for their comments on an earlier draft of this paper.

exact same measure across both waves, regardless of the ever-emergent age, cognitive development, and personal and scholarly experiences of one's test-takers. The measures' content, item wording, response categories, and response formats do not change whatsoever across waves.

Scenario 2: Linkable time-variable measures

Time-variable measures are those whose content, wording, response categories, and/or response formats vary across waves in repeated measures designs. In this scenario, although the time-variable measures are not completely identical across waves, there is at least one anchor item shared by each of the measures, on whose linked (or equated) scores traditional analyses can be performed (Kolen & Brennan, 2004).

Scenario 3: Non-linkable time-variable measures

This scenario involves using measures whose content, item wording, response categories, and/or response formats vary completely across waves. Imagine, for example, a reading achievement test administered at Grade 5 and then Grade 6: The measure administered at Grade 5 cannot be same as that used in Grade 6. If they were the same, the reliability and validity of the test scores would likely be compromised, rendering the study ineffectual (Singer & Willett, 2003). This scenario may also be encountered when one's sample size is small or when one cannot compare the sample's scores to those of a norming group. In such cases, even if the measures share common items, it is not always advisable to link or equate the measures' scores.

Objective

Repeated measures analyses are often characterized by one set of individuals being measured more than once on the same or commensurable dependent variable. Many researchers understand the phrase "same or commensurable dependent variable" to mean that the exact same measure must be used across all waves study.

As Scenario 1 (exact same measure across both waves) illustrates, some constructs can in fact be measured using the exact same measure over time. As Scenario 2 (linkable time-variable measures) and particularly Scenario 3 (non-linkable time-variable measures) describe, however, there are often situations in which one's construct of choice makes the use and re-use of the exact same measure across waves difficult – and even impossible. Seeing as traditional linking/equating techniques are not possible when the measures cannot be made to be identical (Kolen & Brennan, 2004), what is a researcher to do, then, if the use of time-variable measures is necessary?

Therefore, this article focuses on the analysis of two-wave change with linkable – and particularly non-linkable – time-variable measures. Many of the current strategies used to handle time-variable measures (such as vertical scaling and item response theory techniques; see Kolen & Brennan, 2004) are often only useful to large testing organizations that have access to very large numbers of test-takers and expansive item pools, or in those situations in which the time-variable measures share some number of common items. Therefore, the objective of this article is to introduce a workable solution to the problem of analyzing change with time-variable measures administered over two waves – a solution that can be implemented easily in everyday research settings.

The Non-Parametric Difference Score (NPAR-DIFF)

The NPAR-DIFF involves rank transforming or ordering individuals' original test scores within wave, and then using the change (difference) score computed from the respective ranks as the dependent variable in subsequent parametric independent sample *t*-tests. It is this use of ranks, instead of original scores, that makes the NPAR-DIFF a non-parametric solution.

Lloyd (2006) and Lloyd, Zumbo, and Siegel (2007) refer to the general approach of converting original scores into ranks pre-analysis as the Conover solution, in recognition of the influential work of W. J. Conover (e.g., Conover, 1999; Conover & Iman, 1981), whose research not only inspired the NPAR-DIFF, but also provides evidence for the solution's viability.

A rank represents the position of a test-taker on a variable relative to the positions held by all other test-takers on that same variable. Ranking or rank transforming refers to the process of transforming a test-taker's original score to rank relative to other test-takers – suggesting a one-to-one function f from the sample values [e.g.,$\{X_1, X_2, ..., X_N\}$] to the first N positive integers [e.g., $\{1, 2,…, N\}$], (Zimmerman & Zumbo, 1993).

For example, if Test-taker X earned a score of 20 on a given variable, Test-taker Y earned a score of 21, and Test-taker Z earned a score of 22, then the test-takers' respective ranks would be 1, 2, and 3 (where a rank of 1 is given to the test-taker with the lowest score). One may also assign ranks such that the test-taker with the highest score receives a rank of 1; however, it is often easier to think of test-takers receiving the highest score as also receiving the highest rank value.

The NPAR-DIFF's Assumptions

As with all methodological tools, the NPAR-DIFF comes with its own set of assumptions. First, the scales for the measures' original scores must be at least ordinal in nature. Second, the ranks must show heterogeneous change, meaning that all test-takers do not change the same amount across waves (Zumbo, 1999). Imagine that Test-Taker X earns a rank score = 1 across both waves and Test-Taker Y earns a rank score = 2 across both waves. For both test-takers, the change scores computed from the rank equal zero, suggesting homogeneous change – which, for reasons outlined by Zumbo (1999), cannot be used in change analyses. It should be noted that an inability to handle homogeneous change is not a problem endemic to the NPAR-DIFF; homogeneous change also renders ineffectual the calculation of simple difference scores.

Finally, the NPAR-DIFF requires that a commensurable (or comparable or similar) construct is measured across all waves of the study. Commensurability is generally thought to mean that the same primary dimension or latent variable is driving the test-takers' responses at each wave. A latent variable is an unobserved variable that accounts for the correlation among one's observed or manifest variables. In ideal circumstances, measures are designed such that the latent variable that drives test-takers' responses represents the construct of interest.

Example

Suppose a researcher is interested in exploring whether there are gender differences in test-takers' rank-based numeracy assessment difference scores (scores that represent the comparison of test-takers' scores at the second wave of data collection to their respective baseline or initial measure scores). Note that the research question changes slightly when one applies the NPAR-DIFF: No longer are the inferences made from the original scores; rather they are made from the ranks.

To illustrate the implementation of the NPAR-DIFF, Foundation Skills Assessment (FSA) numeracy subtest data from the British Columbia Ministry of Education were obtained. The FSA, an annual assessment administered by the Ministry, is designed to measure the reading comprehension, writing, and numeracy skills of 4th- and 7th-grade students throughout British Columbia. The FSA is administered in public and funded independent schools across the province in late April/early May of each year. Approximately 40,000 students per grade level write the FSA each year.

Obtained was the entire population of standardized numeracy subtest scores of 41,675 test-takers who wrote the FSA in both 1999/2000 (Wave 1, Grade 4) and 2002/2003 (Wave 2, Grade 7). Test-takers who were missing a wave of FSA data were excluded from analyses. Of this population of test-takers, a random 10% convenience sample of 4097 test-takers (n_{female} = 2055; n_{male} = 2042) was retained for analyses. Each test-taker's record included an arbitrary case number, and a gender flag. The Ministry has standardized test-takers' FSA scores such that each wave's score distribution has M = 0 and SD = 1.

Willett, Singer, and Martin (1998) state that standardized test scores should never be used in the place of raw scores in individual growth modeling analyses (readers are referred to their article for the specific reasons why). In this case, however, ranks are being used in the

Table 1

Descriptive Statistics for Each of the Two Waves of FSA Original Scores (N = 4097*)*

Gender	*Original* *Variable Name*	*Min*	*Max*	*M*	*SD*	*Skew*	*Kurtosis*
Female	*grade4original*	-4.83	4.66	-.26	1.10	-.88	4.30
(*n* = 2055)	*grade7original*	-2.08	2.85	.06	.90	.31	-.30
Male	*grade4original*	-4.83	5.33	-.16	1.11	-.78	4.21
(*n* = 2042)	*grade7original*	-2.58	2.85	.11	.92	.31	-.25

place of the original test scores. Thus, it is unimportant whether or not the original test scores come in the form of standardized scores. Furthermore, the Ministry of Education does not supply researchers with raw FSA scores – only standardized scores.

As Table 1 illustrates, the descriptive statistics for each wave of FSA original scores vary across gender and wave. When performing the NPAR-DIFF, data must be entered into the data matrix (spreadsheet) in person-level format, in which one row represents one individual, with time-related variables represented along the horizontal of the spreadsheet (as in Table 2). The key to implementing the NPAR-DIFF is that one first rank transforms the data within wave, with the mean rank being assigned to ties. Table 2 illustrates that Test-Taker *X*, for example, earns a Rank = 2 for Wave 1 (Grade 4) because his original Wave 1 score (0.20) is between those of Test-Taker *Y* (-.15, Rank = 1) and Test-Taker *Z* (1.45, Rank = 3). Recall from an earlier section that a Rank = 1 is assigned to the test-taker with the lowest within-wave score.

Two-Wave Designs: Two Common Change Scores

As described earlier, two-wave designs are characterized by some comparison of an individual's score at the second wave of data scores involved in two-wave designs are:

(a) the simple difference score and
(b) the residualized change score (Zumbo, 1999).

Simple difference score

The most common of all change indices is the simple difference score, which is calculated by simply subtracting a test-taker's score at Wave 1 from his or her score at Wave 2. A positive simple difference score typically indicates an increase over time, whereas a negative score indicates a decrease over time.

Residualized change score

As Zumbo (1999) describes more fully, it has been argued that simple difference scores are unfair because of their base-dependence (i.e., scores at Wave 2 are correlated negatively with scores at Wave 1). As such, the residualized change score was developed as an alternative to the simple difference score. Although there are different ways to create such scores, the most common residualized change score is estimated from the regression analysis of the Wave 2 score on the Wave 1 score. In other words, the estimated Wave 2 score is subtracted from the actual Wave 2 score (whether it be an original or rank).

Table 2. An example person-level data matrix showing two waves of hypothetical original FSA scores and their corresponding within-wave rank scores.

	Example Original Variables		Corresponding Rank Variables	
	grade4original	*grade7original*	*grade4rank*	*grade7rank*
Test-Taker *X*	0.20	0.45	2	1.5
Test-Taker *Y*	-.15	1.35	1	3
Test-Taker *Z*	1.45	0.45	3	1.5

The intrinsic fairness, usefulness, reliability, and validity of the two-wave research design have been debated for decades (Zumbo, 1999). In their seminal article, Cronbach and Furby (1970) disparage the use of two-wave designs, arguing that change scores are rarely useful, no matter how they are adjusted or refined (Cronbach & Furby, 1970). Their disdain of two-wave designs was so strong that they stated that researchers who ask questions using simple difference scores are better advised to frame their questions in other ways (Cronbach & Furby, 1970). As Zumbo (1999) notes, it is somewhat puzzling that there exists the notion that one should avoid two-wave designs at "all costs", given that variations of the difference score lie at the heart of various widely-used and commonly-accepted statistical tests, such as the paired samples t-test.

Determining the Appropriate Change Score to Serve as the Dependent Variable

In order to determine which specific change score should serve as the dependent variable in this particular FSA example, it is necessary to follow the guidelines of Zumbo (1999), who writes that "one should utilize the simple difference score instead of the residualized difference if and only if $\rho(X_1, X2) > \sigma X_1 / \sigma X_2$" (p. 293) – that is, if the correlation between the Wave 1 and 2 scores is greater than the ratio of the respective standard deviations. It is important to stress that, when implementing the NPAR-DIFF solution for two-wave data, one's decision about using the simple difference and residualized change score must be based on test-takers' ranks– not their original scores.

The computed across-gender correlation between the Grade 4 and 7 ranks [$\rho(X_1,X_2)$] was computed as 0.66, compared to 0.99 (1182.84/1182.84) for the ratio of the two standard deviations of rank scores [$\sigma X_1 / \sigma X_2$]. Because the correlation value is less than the ratio value, the rank-based residualized change score is used in the place of the rank-based simple difference score as the dependent variable in the subsequent parametric analysis (Zumbo, 1999).

Explanation of the Statistical Output

A regular independent samples *t*-test was then performed on test-takers' rank-based residualized change scores, with gender identified as the predictor variable. It should be reiterated that the unique aspect of the analysis is that test-takers' rank-based change scores are used in the place of the change scores computed from test-takers' original scores. Original scores are, in a sense, only collected as a means of computing test-takers' ranks. The research question, results, and inferences made from the results must reflect the fact that the scores have been transformed and, hence, the focus is no longer on the original scores.

The independent sample's *t*-test output revealed that the mean rank-based residualized change score for males was -7.64 (*SD* = 882.31) as opposed to 7.59 for females (*SD* = 876.71), meaning that the average Wave 2 rank less the rank at Wave 2 predicted from the Wave 1 rank score is higher for females than for males. This finding suggests that the female test-taker gained 7.5 points in relative standing across the two waves, whereas the average male test-taker's relative standing decreased approximately 7.6 points.

Despite the mean differences in residualized change scores for males and females, the independent samples *t*-test results showed that there is no statistically significant gender difference in the residualized change scores, *t*(4095) = -.555, *p* = .579 (assuming equal variances; two-tailed). Thus, the male test-takers' mean rank-based residualized change score did not differ significantly from that of the female test-takers – suggesting that neither gender's relative standing over time differ significantly from the other.

Even though there was no statistically-significant gender difference found, an effect size was still computed, for reasons outlined by Zumbo and Hubley (1998). A Cohen's *d* effect size was calculated by subtracting the mean residualized change score of one group (females) from that of the other group (males) and dividing that difference by the pooled rank-based standard deviation. The resultant effect size was computed as 0.02, which represents a small effect size (Cohen, 1988).

Strengths of the NPAR-DIFF

The non-parametric difference score, a solution for the problem of analyzing change and growth with time-variable measures collected over two waves is an effective tool for researchers in everyday research settings for the following reasons:

Ease of use

One strength of the NPAR-DIFF is that it is easy to implement. As Conover and Iman (1981) observe, it is often more convenient to use ranks in a parametric statistical program than it is to write a program for a non-parametric analysis. Furthermore, all of the steps required for the implementation of the NPAR-DIFF (i.e., rank transforming data within waves, conducting independent samples t-tests, etc.) can be easily performed using commonly-used statistical software packages.

Marries non-parametric and parametric methods:

Second, by rank transforming the data pre-analysis, parametric and non-parametric statistical methods are combined, providing "a vehicle for presenting both the parametric and nonparametric methods in a unified manner" (Conover & Iman, 1981, p. 128).

Makes use of the ordinal nature of data

Third, the NPAR-DIFF makes use of the ordinal nature of continuous-scored data: A test-taker with a low original score relative to other test-takers in his wave will also yield a low relative rank. Similarly, a test-taker with a high test-score will also yield a high rank. As a result, within-wave order among the test-takers is preserved.

Requires no common/linkable items

Unlike many of the traditional test linking methods and strategies, the NPAR-DIFF can be implemented not only in situations in which one's study involves time-variable measures that can be linked (Scenario 2), but also situations in which the time-variable measures share no linkable items whatsoever (Scenario 3). Hence, unlike vertical scaling, equating, and their linking counterparts, the NPAR-DIFF provides a means by which researchers can study change – whether or not the measures contain linkable items.

Requires no norming group

Due to time and financial constraints, it is not always possible to compare the scores of one's sample to those of an external norming sample. As such, an additional strength of the NPAR-DIFF is that it can be conducted using simply the scores of the sample of test-takers, thereby eliminating the need for a group to which to compare the sample's scores.

Limitations of the NPAR-DIFF

As with any methodological tool, the NPAR-DIFF has various limitations. Within-wave ranks are bounded. Rank transforming refers to the process of converting a test-taker's original score to rank relative to other test-takers. The values assigned by the function to each sample value in its domain are the number of sample values having lesser or equal magnitude. Consequently, the ranks are bounded from above by *N*. As a result, "any outliers among the original sample values are not represented by deviant values in the rank" (Zimmerman & Zumbo, 1993, p. 487).

Suppose on a standardized test of intelligence, Test-Taker *W* earns a score 100, Test-Taker *X* earns a score of 101, Test-Taker *Y* earns a score of 102, and Test-Taker Z earns a score of 167. Test-Taker *Z*'s score, relative to the other test-takers, is exceptional. Despite the exceptional performance on the measure, the test score is masked by the application of ranks: Test-taker $W = 1$, Test-taker $X = 2$, Test-taker $Y = 3$, and Test-taker $Z = 4$.

As a result, one limitation of the NPAR-DIFF is that there may be problems associated with the inherent restriction of range it places on data. Differences between any two ranks range between 1 and $N - 1$, whereas the differences between original sample values range between 0 and infinity (Zimmerman & Zumbo, 1993).

Difficulties associated with handling missing data

Recall that only those test-takers for whom data were available at both waves were retained in the analyses. As most educational, health, and social science researchers will agree, no discussion about change and growth is complete without a complementary discussion about one unavoidable problem: missing data. In longitudinal designs, particularly those that span months or years, it is extremely common to face problems associated with participant dropout, attrition, and as well as participants who join, or return to the study, in later waves.

One possible strategy for circumventing, or at least mitigating the effect of, missing data is to impute the missing original scores prior to rank-transforming the data within-wave pre-analysis. Schumacker and Lomax (2004) discuss various missing data imputation methods.

Makes use of the ordinal nature of data:

Recall that the fact that the NPAR-DIFF makes use of the ordinal nature of continuous-scored data was previously identified as one of the solution's strengths. As Lloyd (2006) and Lloyd, Zumbo, and Siegel (2007) observe, precisely what the NPAR-DIFF wins by, it also loses by: Because of the rank transformation of the original scores, differences between raw scores are not necessarily preserved by the corresponding ranks. For example, a difference between the raw scores corresponding to the 15th and the 16th ranks is not necessarily the same as the difference between the raw scores corresponding to the 61st and 62nd ranks in a collection of 500 test scores (Zimmerman & Zumbo, 2005, p. 618).

Conclusion

Investigating the problem of analyzing change and growth with time-variable measures is important for two reasons. First, as Willett et al. (1998) and von Davier, Holland, and Thayer (2004) describe, the rules about which tests are permissible for repeated measures designs are precise and strict. Given these conditions, it is necessary to investigate how repeated measures analyses can be made possible – psychometrically and practically – when the measures themselves change across waves.

Second, given the sizeable growth in longitudinal large-scale testing in recent years, it is necessary to find a viable and coherent solution to the problem so that researchers can make the most accurate inferences possible about their test scores.

Recognizing the importance of this problem, this article introduced a workable solution for handling the analysis of change over two waves, when the measures used at each wave are not the same. Although useful in many research settings, the non-parametric difference score (NPAR-DIFF) is by no means a universal panacea and should, therefore, be used judiciously and in accordance with the aforementioned assumptions. Given that the problem of time-variable measures has, to date, gone relatively unaddressed in the change/growth and test linking literatures, it is imperative that future research explores this profoundly important, problem to a much fuller degree.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum, Associates.

Conover, W. J. (1999). *Practical nonparametric statistics (3rd ed.).* New York: John Wiley & Sons.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" - Or should we? *Psychological Bulletin, 74,* 68-80.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.).* New York: Springer-Verlag.

Lloyd, J. E. V. (2006). *On modeling change and growth when the measures themselves change across waves: Methodological and measurement issues and a novel non-parametric solution.* Unpublished doctoral dissertation, University of British Columbia.

Lloyd, J. E. V., Zumbo, B. D., & Siegel, L. S. (2006). *The non-parametric HLM: A workable solution for analyzing change and growth when the measures themselves change across waves.* Manuscript submitted for publication.

Schumacker, R. E, & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling, 2^{nd} edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford Press.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of equating.* New York: Springer.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10,* 395-426.

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences, Volume 1: Methodological issues* (pp. 481-517). Hillsdale, NJ: Lawrence Erlbaum.

Zimmerman, D. W., & Zumbo, B. D. (2005). Can percentiles replace raw scores in statistical analysis of test data? *Educational and Psychological Measurement, 65,* 616-638.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In Bruce Thompson (Ed.). *Advances in Social Science Methodology, Volume 5,* (pp. 269-304). Greenwich, CT: JAI Press.

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *Journal of the Royal Statistical Society, Series D (The Statistician), 47,* 385-388.